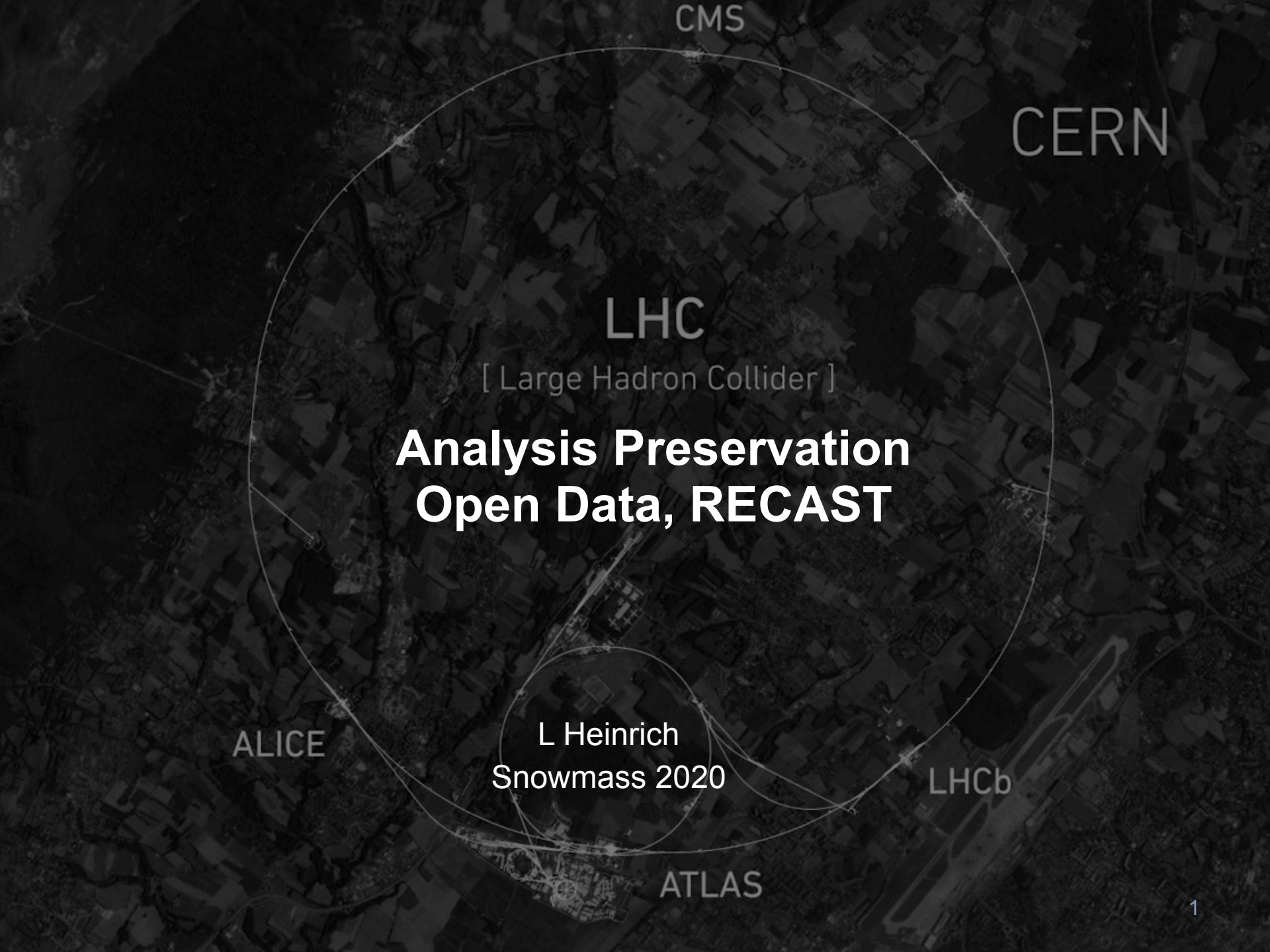
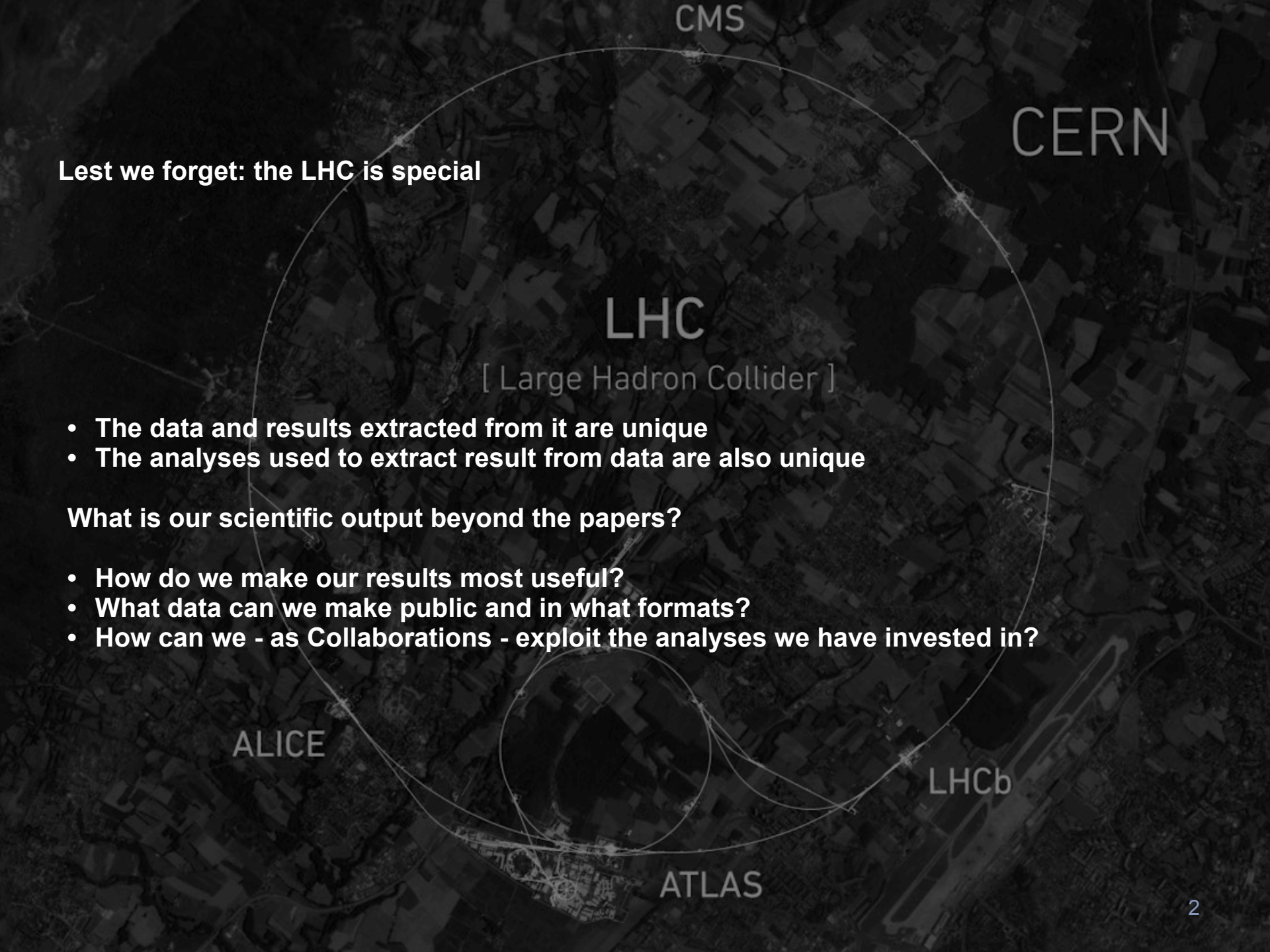


Analysis Preservation
Open Data, RECAST

L Heinrich
Snowmass 2020



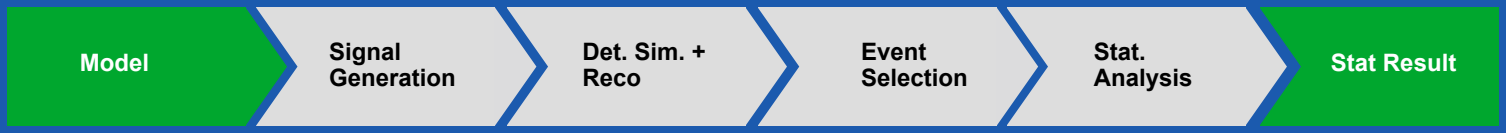
Lest we forget: the LHC is special

- **The data and results extracted from it are unique**
- **The analyses used to extract result from data are also unique**

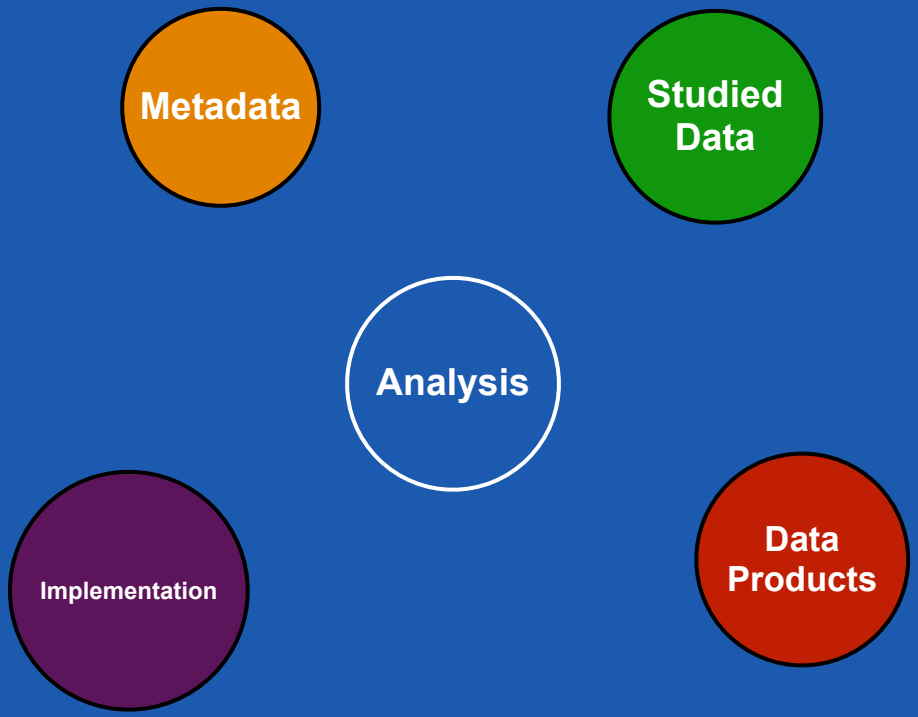
What is our scientific output beyond the papers?

- **How do we make our results most useful?**
- **What data can we make public and in what formats?**
- **How can we - as Collaborations - exploit the analyses we have invested in?**

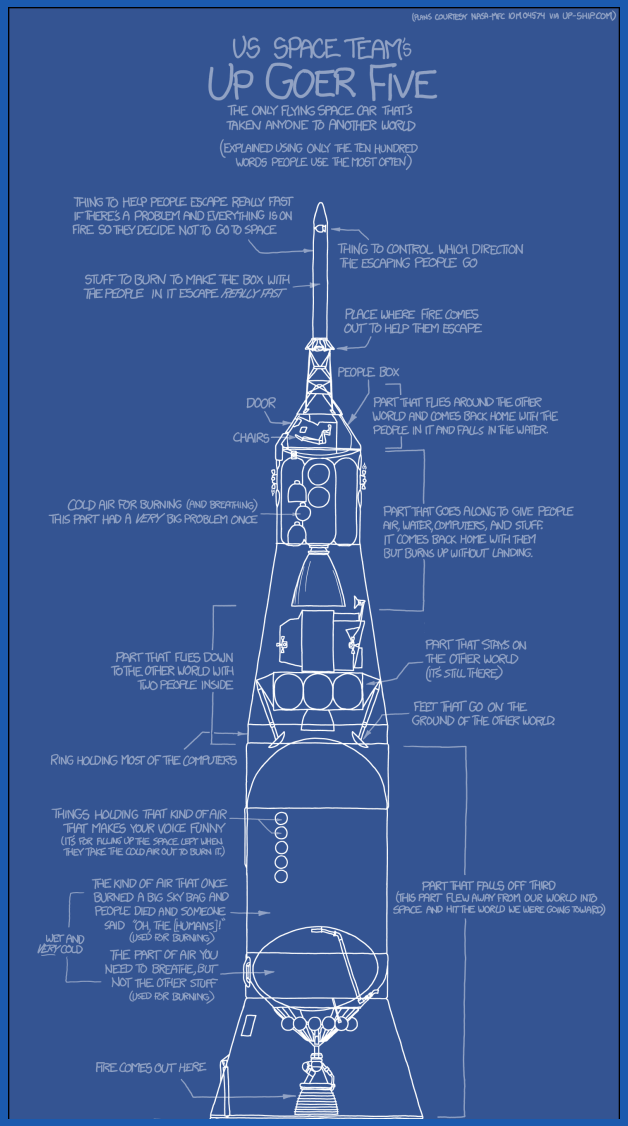
Analysis Sketch



Preservation Domains



External

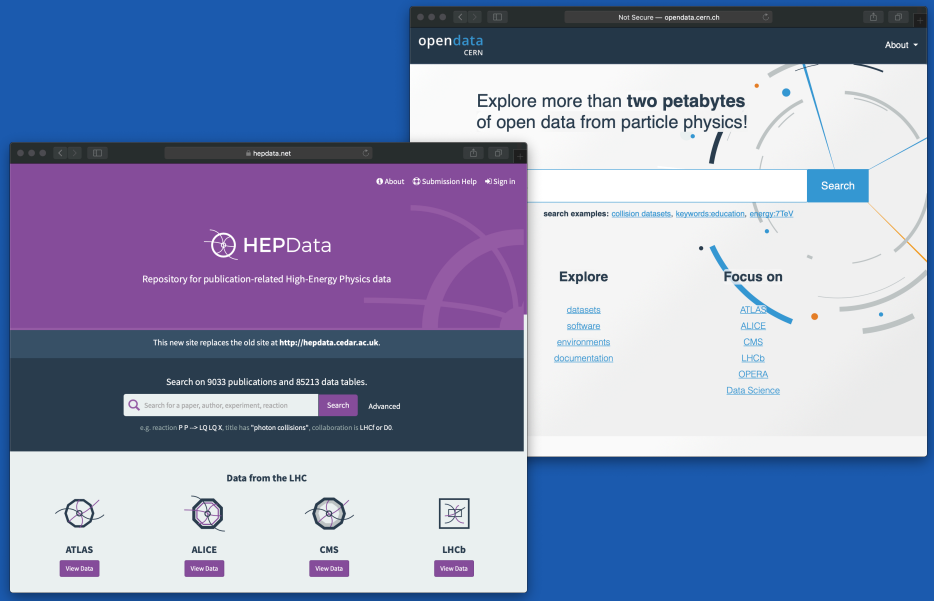


Internal

CERN

Three broad areas of activity

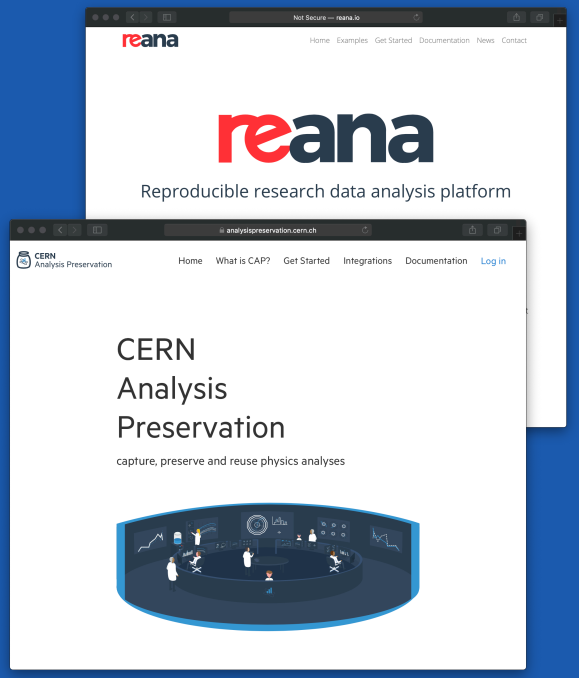
External



Analysis Data Products and Result Preservation

Open Data for Outreach, Education and Research

Internal



Reproducible Workflows & Analysis Preservation

HepData has been the main vehicle to provide

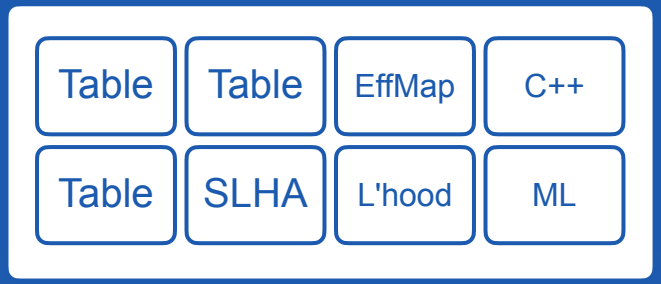
high quality public data products for published analyses

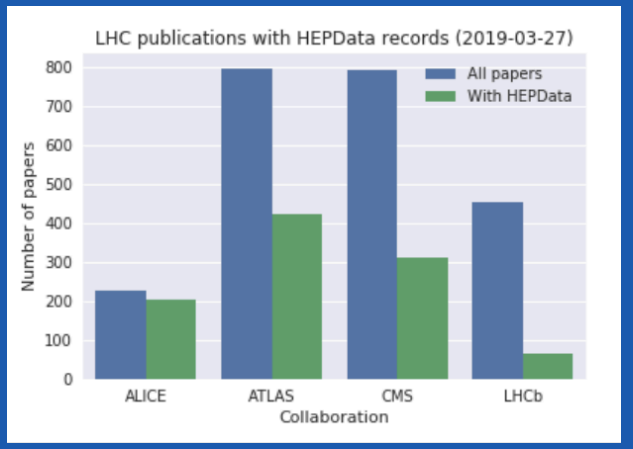
publicly available. All LHC experiments rely on this.

- **HepData submission often required for analysis approval**

Types of data products expanded from

to broader collection of data





ALICE: 90%
ATLAS: 52%
CMS: 39%
LHCb: 14%

Additional Material helps approximate reimplementation of data analyses w/ e.g. Rivet (can cover also BSM and HI)

```
#include "SimpleAnalysis/AnalysisClass.h"
#include "SimpleAnalysis/NtupleMaker.h"
#include "SimpleAnalysis/PDFReweight.h"
#include <LHAPDF/LHAPDF.h>
#include "TMath.h"

DefineAnalysis(EwkOneLeptonTwoBjets2018)
// Wh->1l+bb+met analysis (Run2 data)

void EwkOneLeptonTwoBjets2018::Init()

{
  // Define signal/control regions
```

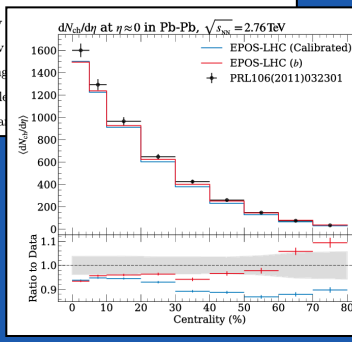
```
// -*- C++ -*-
#include "Rivet/Analysis.hh"
#include "Rivet/Projections/ChargedFinalState.hh"
#include "Rivet/Tools/Correlators.hh"
#include "Rivet/Tools/AliceCommon.hh"
#include "Rivet/Projections/AliceCommon.hh"

namespace Rivet {


  /// @brief Multiparticle azimuthal correlations pp, pPb, XeXe and Pb
  class ALICE_2019_I1723697 : public CumulantAnalysis {
  public:

    /// Constructor
    ALICE_2019_I1723697() :
      CumulantAnalysis("ALICE_2019_I1723697") {}
```

LU TP 20-04
MCNET-20-04

Confronting Experimental Data with Heavy-Ion Models

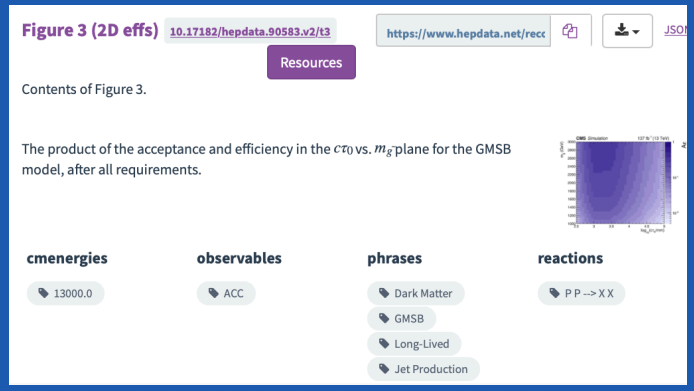
RIVET for Heavy Ions

Christian Bierlich,[1,2] Andy
Peter Harald Lindenov
Jan Fiete Grosse-Oetring
Patrick Kirchgaeßer,[6] Jochen Kl
Christine O. Rasmussen,[2] Ma



C++ Code Snippets as starting point, or (better) full Rivet Routine

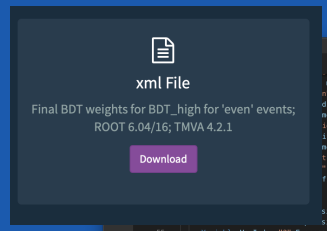
Efficiency Maps:

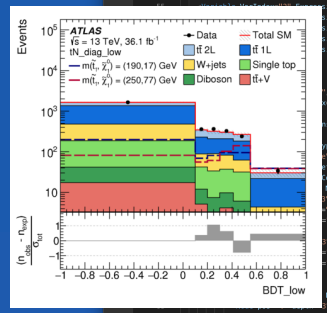


[CMS Record]

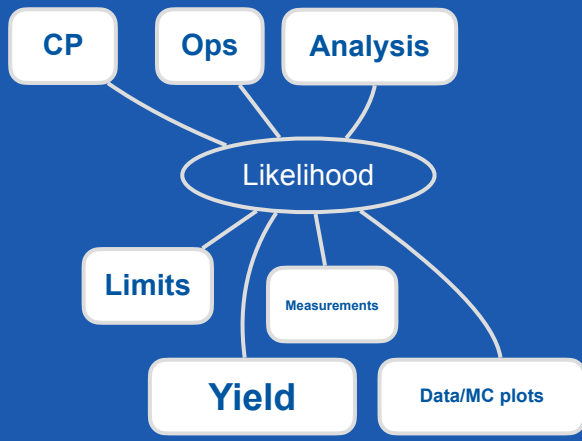
ML models uploaded to HepData

[ATLAS Record]





A new Frontier: Public Likelihoods



$$p(\text{theory}|\text{data}) \sim p(\text{data}|\text{theory}) \cdot p(\text{theory})$$

likelihood: experimentalists

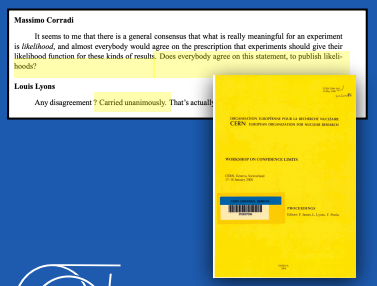
prior: theorists

The likelihood is the central object in analysis

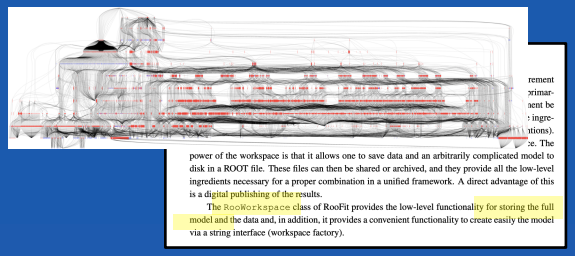
- **the best data product we can provide in principle**

Often HepData information (yields, uncertainties...) is used to <u>reconstruct approximate likelihood</u>

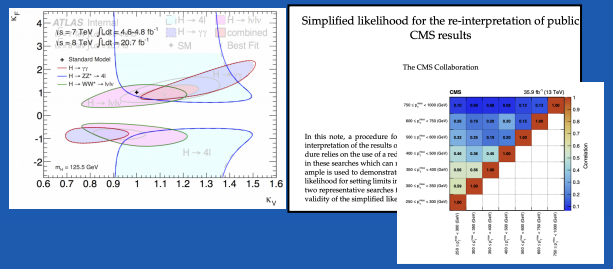
2000 2010 2012 2017 2019



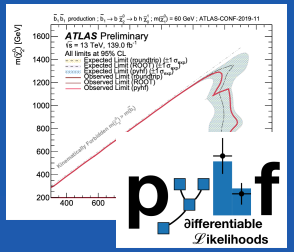
Early Intent



Technical Foundation (Workspaces)



Approximate Realizations



First full releases

HepData is for experimentalists interacting with wider community by releasing public information about already existing data analyses.

But it may not be enough to interact with community on developments of new analyses techniques.

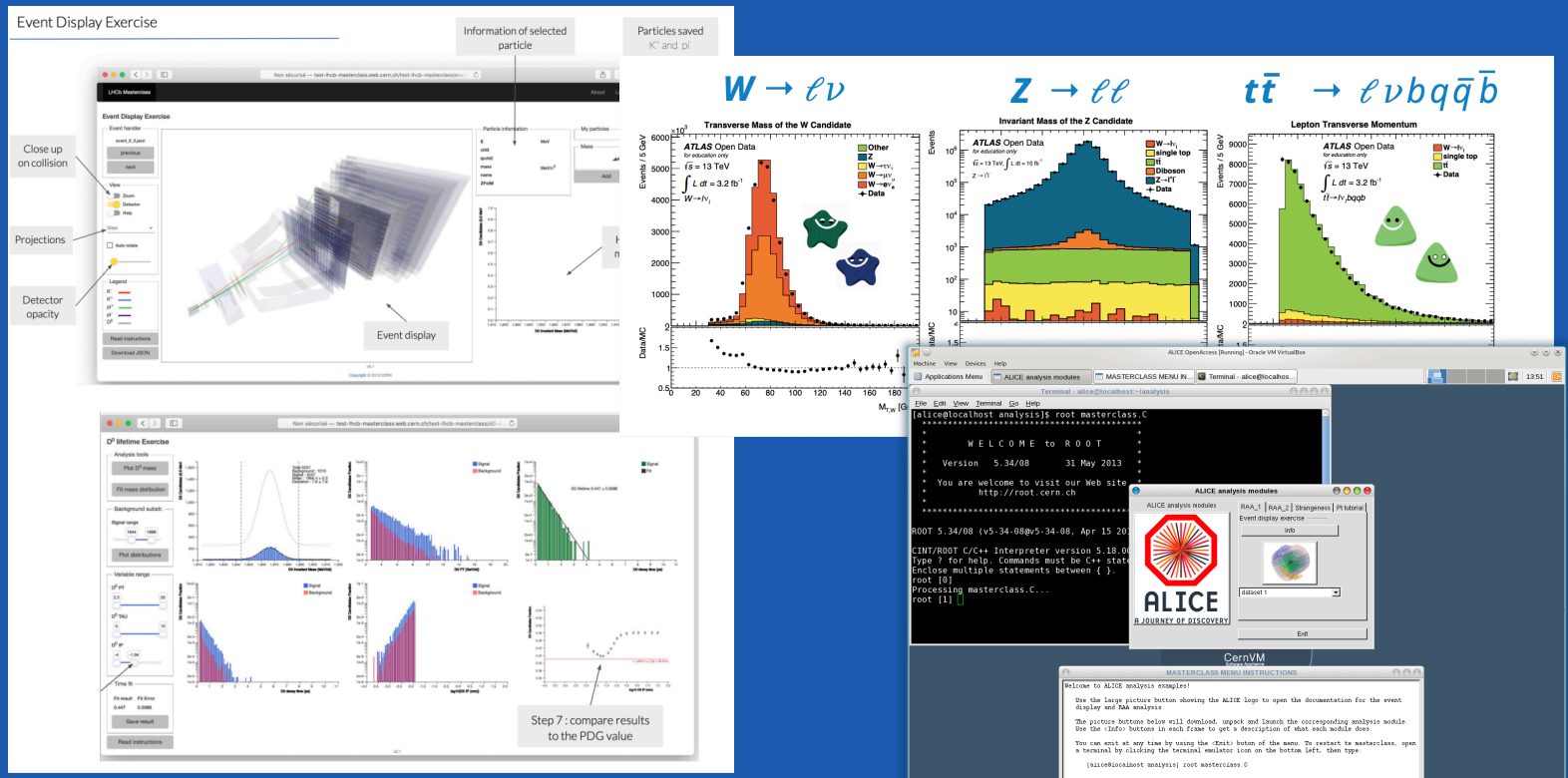
For this we might require a more free-form mode of collaboration: Open Data

CERN

Open Data

All LHC Experiments have Open Data Programs

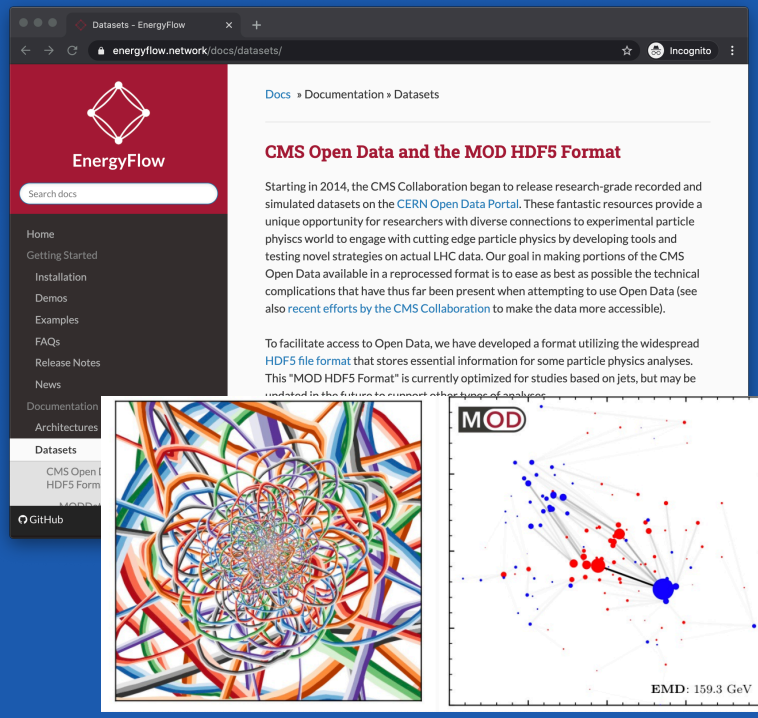
- integrated into CERN Open Data Portal
- ATLAS, LHCb, ALICE so far focused mainly on Outreach & Education



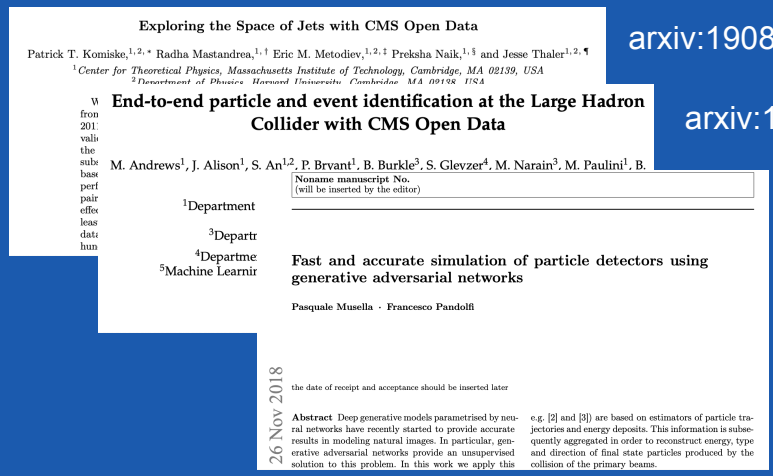
CMS has more expansive Open Data Program for Research

We see external eco-system developing

- Workshop in October [link]



Number of Papers appearing on e.g. Machine Learning methods for LHC


Exploring the Space of Jets with CMS Open Data

Patrick T. Komiske,[1,2,*] Radha Mastandrea,[1,†] Eric M. Metodiev,[1,2,‡] Preksha Naik,[1,§] and Jesse Thaler[1,2,¶]

[1]Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2]Department of Physics, Harvard University, Cambridge, MA 02138, USA

arxiv:1908.08542

End-to-end particle and event identification at the Large Hadron Collider with CMS Open Data

M. Andrews[1], J. Alison[1], S. An[1,2], P. Bryant[1], B. Burkle[3], S. Glevzer[4], M. Narain[3], M. Paulini[1], B.

[1]Department

[3]Departr

[4]Departme

[5]Machine Learnir

arxiv:1910.07029`

Noname manuscript No.
(will be inserted by the editor)

Fast and accurate simulation of particle detectors using generative adversarial networks

Pasquale Musella · Francesco Pandolfi

the date of receipt and acceptance should be inserted later

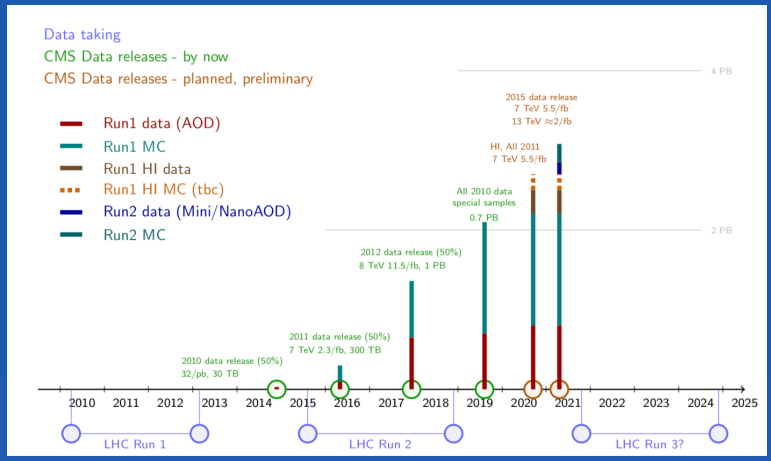
Abstract Deep generative models parametrised by neural networks have recently started to provide accurate results in modeling natural images. In particular, generative adversarial networks provide an unsupervised solution to this problem. In this work we apply this

e.g. [2] and [3]) are based on estimators of particle trajectories and energy deposits. This information is subsequently aggregated in order to reconstruct energy, type and direction of final state particles produced by the collision of the primary beams.

26 Nov 2018

arxiv:1805.00850`

Release Schedule being finalized for coming years



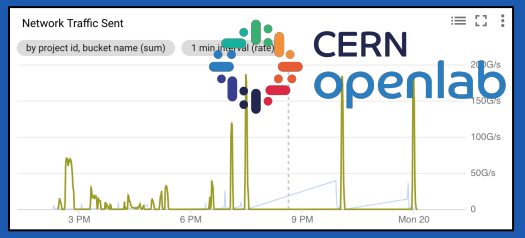
Currently we see a range in approaches. Questions raised by Open Data (L3):

- Can we ensure ability to perform analysis of sufficient quality
- What data formats / software would be released? What's the level of support (if any?)
  - how "final" should objects be (ability to re-reconstruct...)

- Protect Collaboration / Cohesion
  - without a strong collaboration preparation of high-quality OD impossible

- Is analyzing PB scale data really feasible for external users.
  - emerging public cloud infrastructure might help provide on-demand access to scale

  - maybe targeted datasets (e.g. for ML R&D) rather than blanket Open Data?



KubeCon 2019 Keynote [link]

CERN is seeking to harmonize them via a common Open Data Working Group.

HepData is for experimentalists interacting with wider community by releasing public information about already existing data analyses.

Open Data might be useful for experimentalists to interact with external researches on new R&D

But both approaches are not enough.

HepData is analysis-specific, but lossy.
Open Data for new work outside of expt's

→ need infrastructure for internal, lossless analysis preservation

CERN

Internal analysis preservation can capture detail unavailable in other modes of data/analysis preservation.

Increasing complexity in analyses to fully exploit potential of LHC dataset

- **low-level observables: (e.g. BDT on calorimeter clusters)**
  - **even if we publish BDT / NN weights, can you reliably simulate those low-level details?**
- **whole-event observables (NN inputs from many objects)**
  - **simple description of signal model acceptance via e.g. efficiency tables will not work anymore**
- **Need a way to peserve analyses part of result pipeline at full fidelity.**

Internal Reuse:

Efforts by all LHC experiments to foster internal analysis preservation. Ingredients for AP:

capture software

archive analysis code incl. dependencies

capture commands

what do with the captured software

capture workflow

order of individual steps

data assets

input data needed to run the analysis

CERN provides infrastructure to assist experiments

REANA: workflows-as-a-service

CAP: store workflow and other analysis artifacts (software, etc)

reana

Reproducible research data analysis platform

CERN
Analysis Preservation

capture, preserve and reuse physics analyses

CERN

1. **capture software**

   archive analysis code incl. deps.

2. **capture commands**

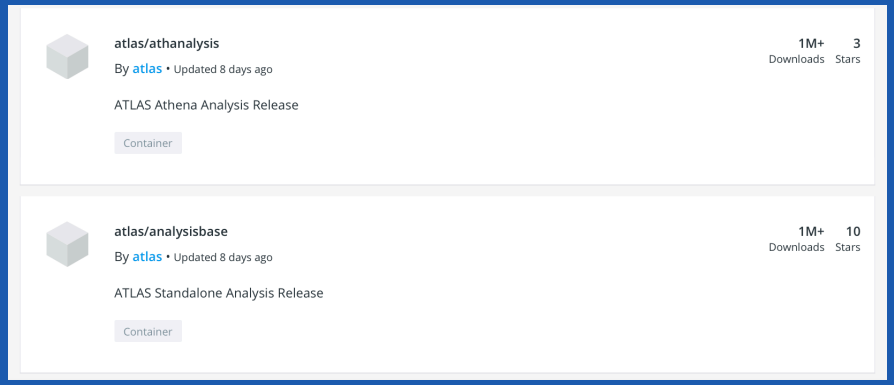
   what do with the captured software

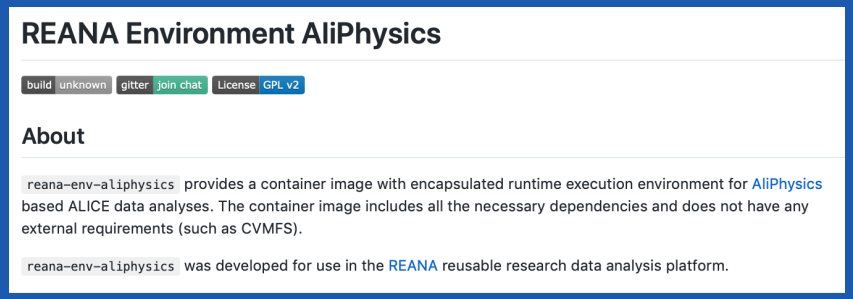
3. **capture workflow**

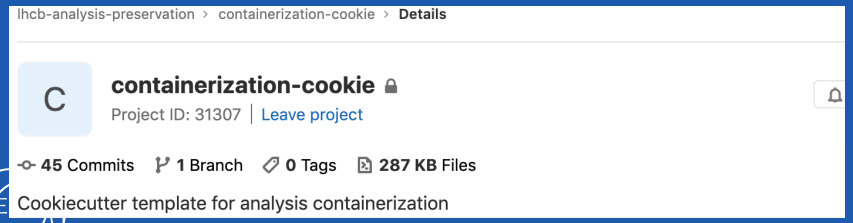
   order of individual steps

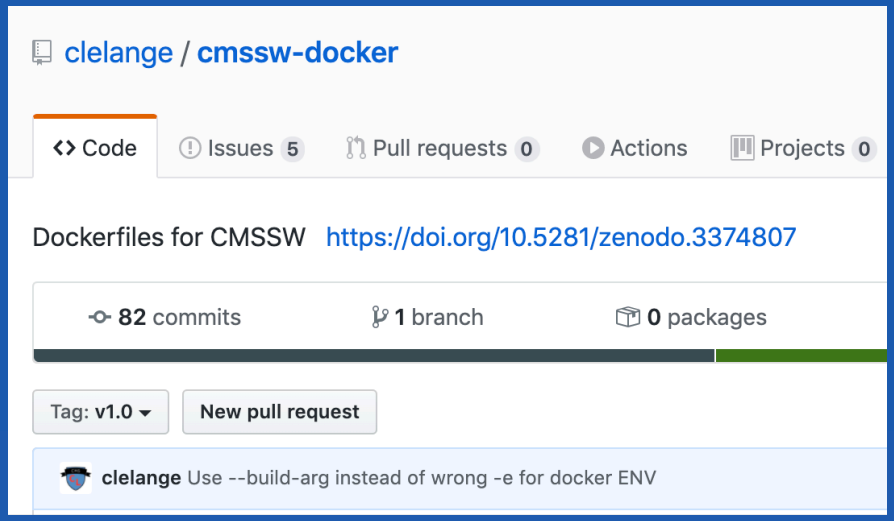
Containers universally seen as suitable techology:

all experiments have some infrastructure to run experiment / analysis code in containers









1. **capture software**

   archive analysis code incl. deps.

2. **capture commands**

   what do with the captured software

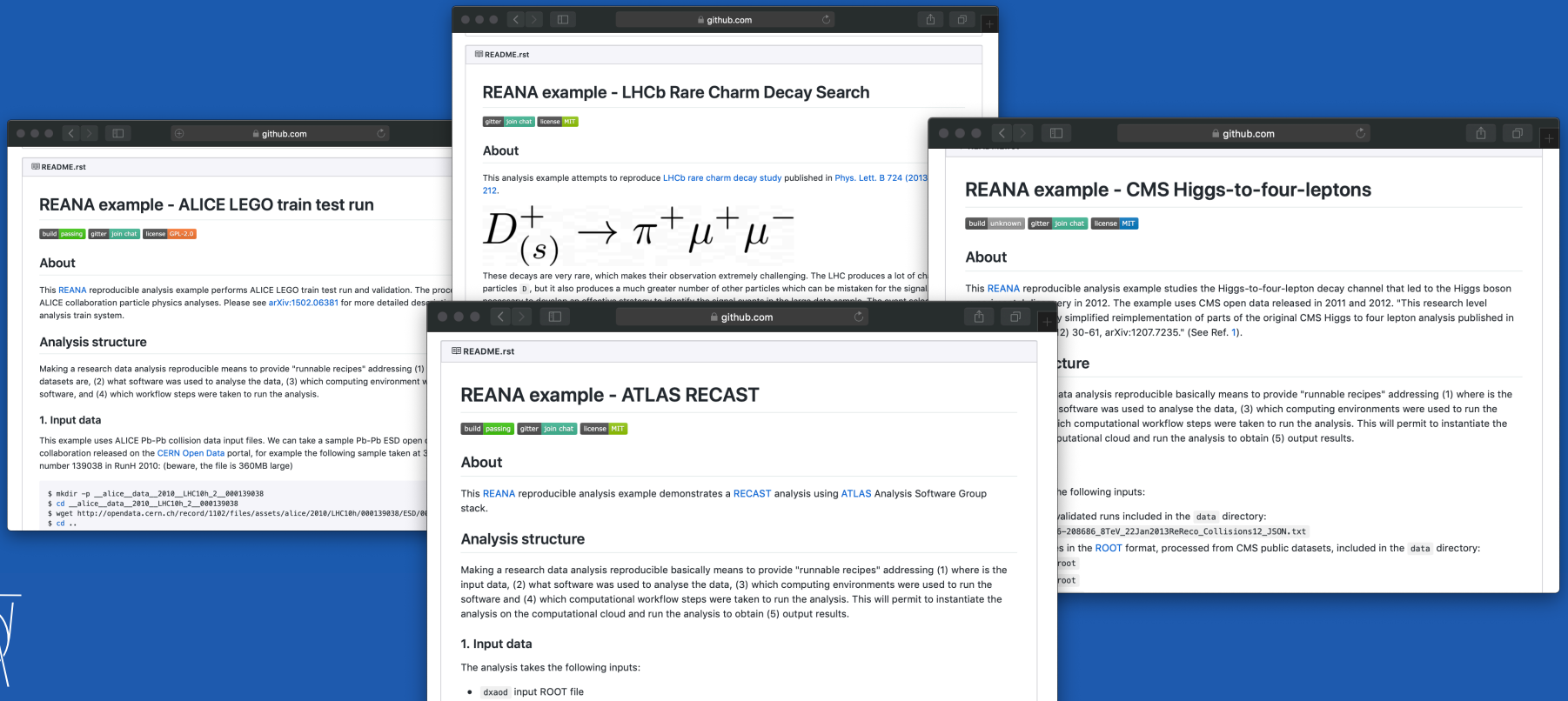
3. **capture workflow**

   order of individual steps

Workflow languages seem to be a good choice:

REANA supports Common Workflow Language, Yadage

- looking into snakemake (LHCb also has snakemake starter kit)

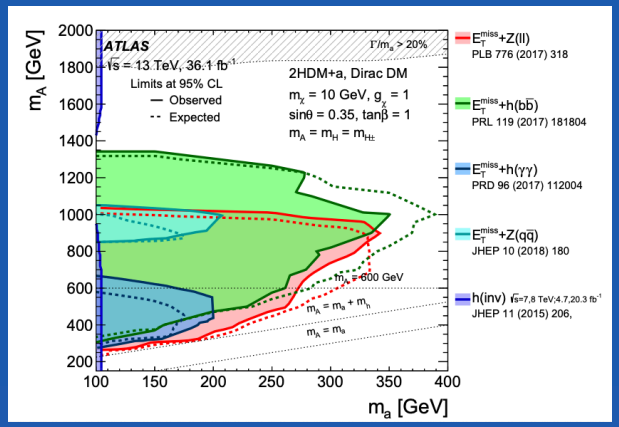


Major use-case for internal re-use: reinterpretation

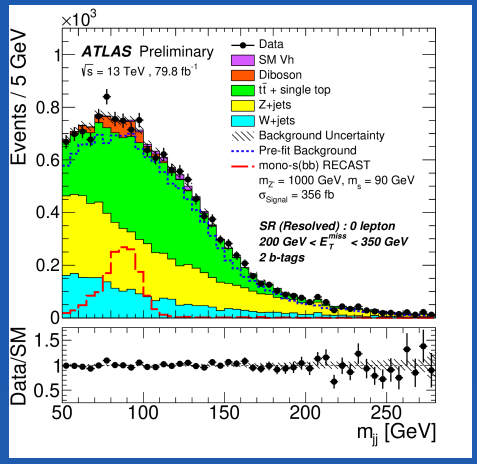
original model alternative model

ATLAS: require analyzers to preserve analysis that at least reinterpretation w/ REANA is possible → realization of RECAST (docker images, scripts, workflows)

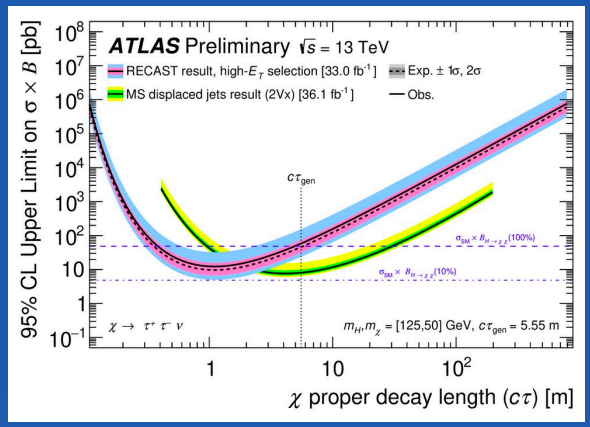
New scientific results based on this (rather technical) requirement



arxiv:1903.01400

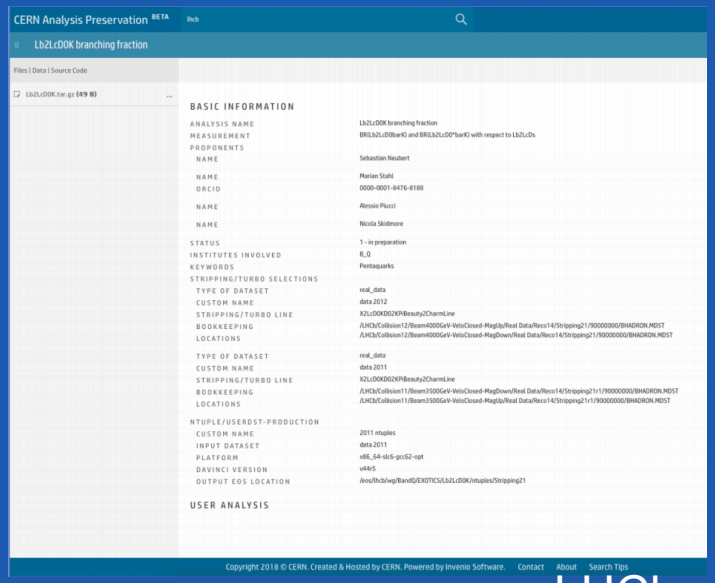


ATL-PHYS-PUB-2019-032

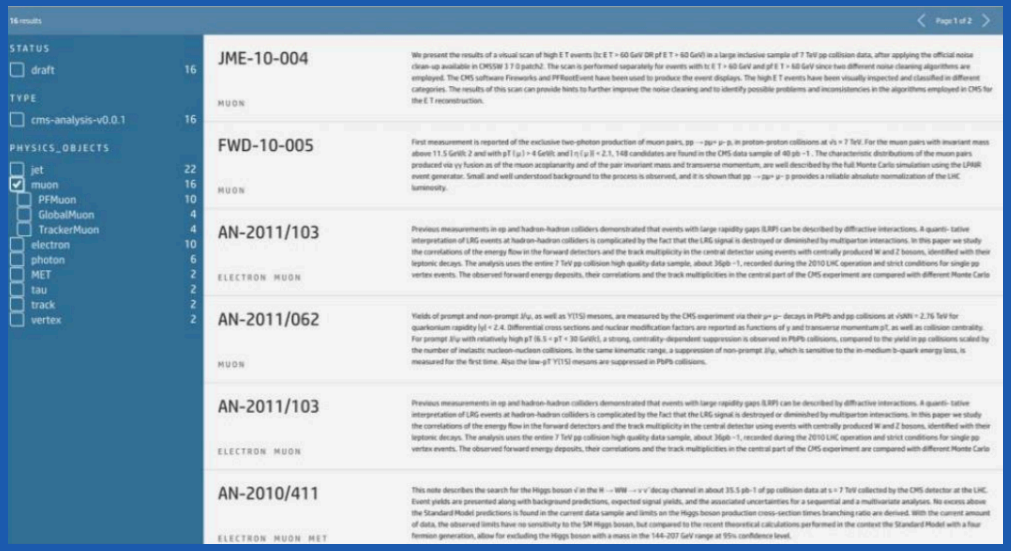


ATL-PHYS-PUB-2020-007

CERN Analysis Preservation Examples



LHCb



CMS

Focus on:

- **ease of use for analysis teams**
  e.g. auto-complete, automatic ingestion, command line clients

- **ease of use for users**
  - **discoverability / search**
  - **integration with REANA**

We see a pattern:

Likelihood: Preserving general likelihood functions in a ssustainable way is a hard problem.

→ restricting to binned models (HistFactory) enabled progresss by narrowing scope

"When solving a problem of interest, do not solve a more general problem as an intermediate step"
- V. Vapnik

AP: Preserving Analyses in full generality (does it work in 100 years?) is too big a problem. RECAST focuses on

- **near-/mid-term solution (e.g. assume containers)**
- **subset of the problem of reinterpretation, not e.g. re-estimation of background)**

CERN

We seem to be entering a golden era of data / analysis preservation for LHC.

Absence of new physics (so far!) forces us to focus on full exploitation of data.

We finally have the technology:

- **containers / workflows for software/analysis preservation**
- **at-scale compute on-demand for Open Data**

See some sociological shifts in community:
- **likelihood releases**
- **analysis preservation (RECAST) as approval requirement**

Good opportunity for new strategic efforts for both open and internal preservation efforts.

CERN